NRL Report 8043

600-Bit-Per-Second Voice Digitizer (Linear Predictive Formant Vocoder)

GEORGE S. KANG AND DAVID C. COULTER

Systems Integration and Instrumentation Branch
Communications Sciences Divison

November 4, 1976

NAVAL RESEARCH LABORATORY
Washington, D.C.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NRL Report 8043	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE *(and Subtitle)*: 600-BIT-PER-SECOND VOICE DIGITIZER (LINEAR PREDICTIVE FORMANT VOCODER)	5. TYPE OF REPORT & PERIOD COVERED: Interim report on a continuing NRL problem
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): George S. Kang and David C. Coulter	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Naval Research Laboratory, Washington, D.C. 20375	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NRL Problem R01-62, Project 33401N, Task XCC51
11. CONTROLLING OFFICE NAME AND ADDRESS: Department of the Navy, Naval Electronic Systems Command, Washington, D.C. 20360	12. REPORT DATE: November 4, 1976
	13. NUMBER OF PAGES: 37
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*	15. SECURITY CLASS. *(of this report)*: Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Voice Digitizer
Formant tracking
Pattern matching
Linear predictive encoding

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Past efforts to achieve practical formant vocoders have been plagued with problems of formant tracker instability, resulting in unnatural "warbles" in the synthesized speech. A new approach to formant frequency determination, combined with a digital implementation, promises to eliminate these effects and to yield a useful formant vocoder. Additional redundancy reduction of information is obtained by means of a pattern-matching technique, which encodes the three formant frequencies into seven bits per frame to provide speech synthesis at 600 bps.

(continued)

20. Abstract (Continued)

This is accomplished by using an existing 2400-bps linear-predictive-encoder (LPE) with some additional processing. A demonstration record of processed speech at 600 bps is included with the report.

CONTENTS

600-BIT-PER-SECOND VOICE DIGITIZER (LINEAR PREDICTIVE FORMANT VOCODER)

INTRODUCTION

This report presents an analysis/synthesis method whereby speech may be transmitted at 600-bits-per-second (bps), a data rate which is less than 1 percent of the pulse-code-modulation (PCM) transmission rate for original speech sounds. This R&D effort was motivated by the pressing need for very-low-data-rate (VLDR) voice digitizers to meet some of the present Navy voice communication requirements. The use of a VLDR voice digitizer makes it possible to transmit speech signals over adverse channels which support data rates of only a few hundred bps or to transmit speech signals over more favorable channels with redundancies for error protection or for other useful applications. The 600-bps synthesized speech loses some of its original speech quality, but the intelligibility is sufficiently high to permit the use of the system in certain specialized military applications.

One of the most attractive features of the VLDR voice-digitizer technique presented in this report is that it is a simple extension of a 2400-bps linear predictive encoder (LPE) (Fig. 1) which has been under intensive investigation by the Navy and other various government agencies and is presently entering advanced development. It is anticipated that

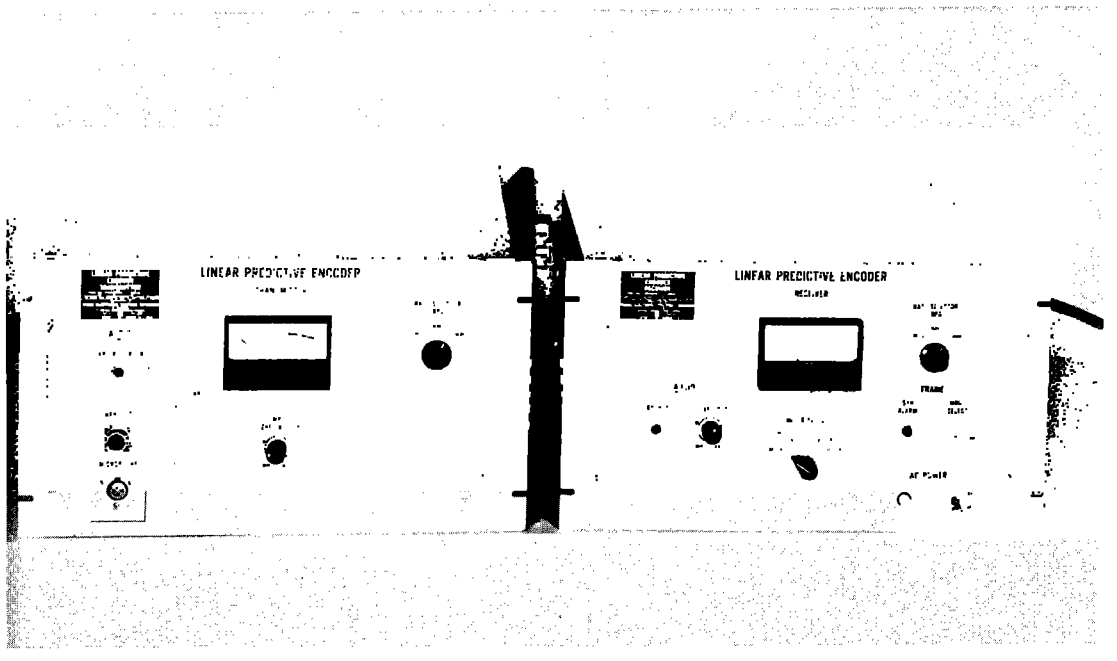


Fig. 1 — Navy experimental 2400-bps linear predictive encoder (LPE)

2400-bps LPEs will be extensively deployed in support of DOD and other government communications. In essence the 600-bps voice digitizer is a 2400-bps LPE with an add-on processor at the transmitter and the receiver. This add-on processor converts the 2400-bps speech data to 600-bps speech data at the transmitter and reconverts the data to 2400 bps at the receiver.

To elaborate this point, the parameters encoded by a typical 2400-bps LPE will be briefly reviewed. An LPE derives two sets of parameters from speech waveforms. One is a set of predictive coefficients, estimated by the least-squares method, which describes the signal transformation characteristics of the vocal tract. The other set describes the excitation waveforms, i.e., pitch period, power level, and voice/unvoice decision (buzz/hiss selection), that define the driving signal for the vocal-tract filter. The vocal-tract filter is a digital recursive filter in which filter parameters are predictive coefficients. In a typical 2400-bps LPE all of these parameters are derived once every 22.5 milliseconds and quantized in 54 bits.

With respect to the parameters transmitted by a 2400-bps LPE, the following modifications are incorporated in the 600-bps voice digitizer:

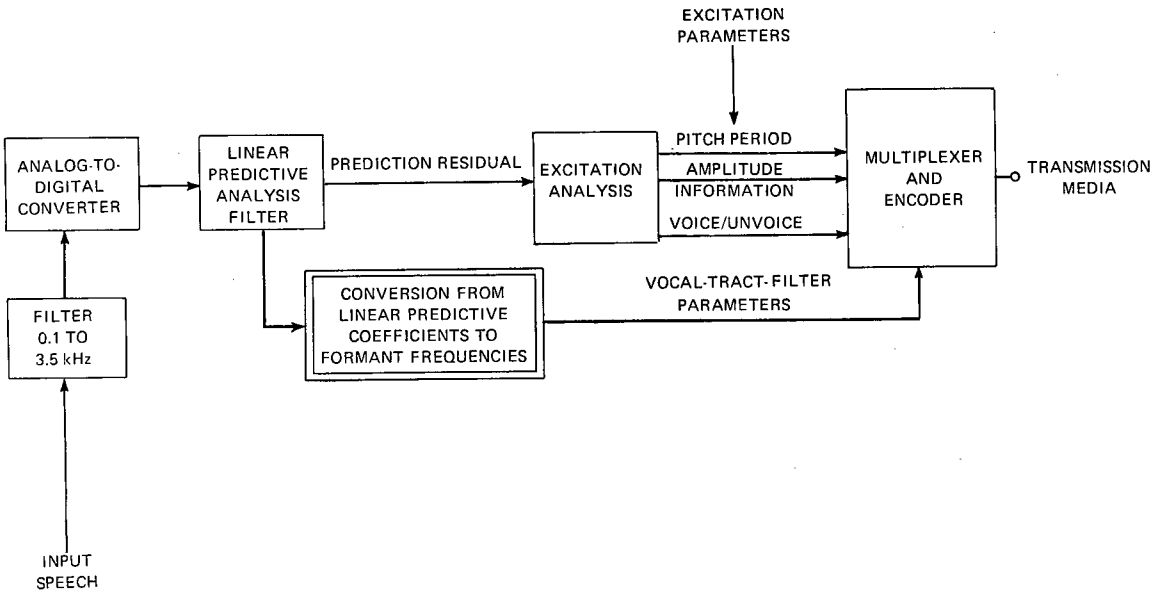
- The parameter update interval is increased from 22.5 to 25 milliseconds. This 10-percent increase is an undesirable but necessary compromise between the 2400-bps LPE updata rate and the number of bits per frame to realize an overall data rate of 600 bps.

- The excitation parameters are virtually identical to those for a 2400-bps LPE, but the pitch period is updated once every other frame. Transmission of the pitch period and the excitation power level require 260 bps, or 43 percent of the overall data rate. This high percentage of the transmission rate is considered necessary, since the pitch period and the excitation power level are essential for natural speech reproduction, which is so vital to acceptable voice communications.

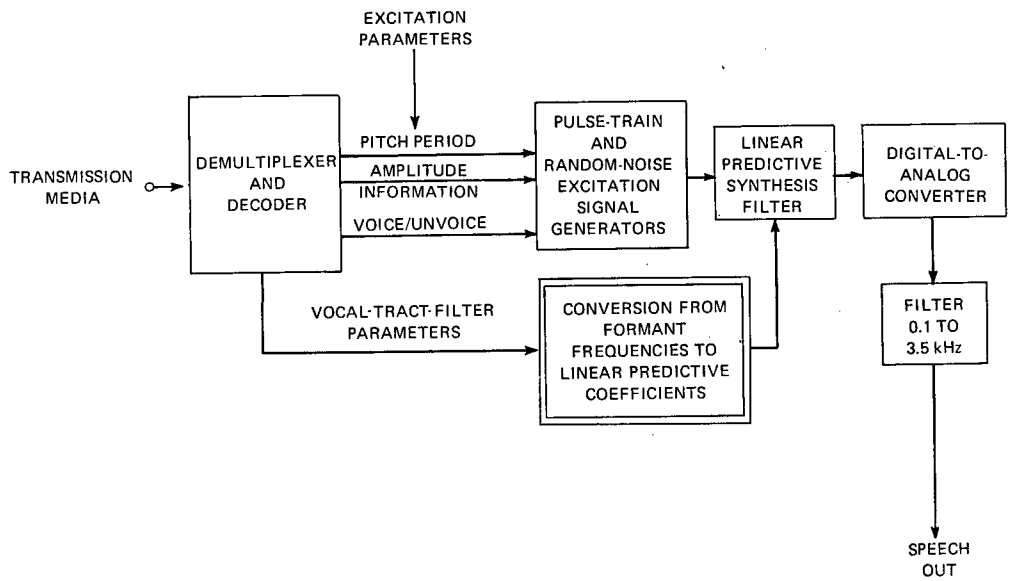
- The vocal-tract-filter parameters take two forms depending on the voicing state: the formant frequencies for voiced sounds, and predictive coefficients for unvoiced sounds. Voiced sounds (mostly vowels) are well characterized by the impulse response of the vocal-tract filter having three resonance frequencies (the first three formant frequencies). Therefore, if speech is voiced, three formant frequencies, derived from predictive coefficients, are transmitted. To economize the data rate, neither the formant bandwidths nor the formant intensities are transmitted. On the other hand, predictive coefficients are directly transmitted for unvoiced sounds (fricatives), because they are poorly characterized in terms of formant frequencies.

Figure 2 is a block diagram of the 600-bps voice digitizer. The blocks bordered by the double lines indicate the processors added to a 2400-bps LPE to provide a 600-bps transmission capability.

The most critical process in the 600-bps voice digitizer is formant tracking. The majority of previous formant-tracking methods relied on some form of spectral analysis of the speech waveform, which is in essence the evaluation of the vocal-tract-filter transfer function along the unit circle in the z plane. Although the spectral analysis is relatively



(a) Transmitter



(b) Receiver

Fig. 2 — 600-bps voice digitizer

simple, it is often unable to detect the vocal-tract-filter poles located well inside the unit circle; i.e., the frequency spectrum does not peak sharply at the frequencies corresponding to the arguments of poles. Likewise two sets of closely adjacent poles are often detected as one pole, leading to a formant-misidentification problem. Such phenomena are commonly observable in speech spectrographs. Thus it is not surprising that formant tracking has long been regarded as impractical (the Background section will give additional information).

The 600-bps voice digitizer described in this report uses predictive coefficients for formant tracking. The use of predictive coefficients as source material for formant tracking has merit because the coefficients appear in the expression of the vocal-tract-filter transfer function as a simple algebraic form:

$$H_n(z) = \frac{1}{1 - \alpha_{1|n}z^{-1} - \alpha_{2|n}z^{-2} - \ldots - \alpha_{n|n}z^{-n}},$$

where the α's are predictive coefficients, z is a complex variable, and n is the total number of filter coefficients (the order of prediction). The roots of the denominator provide the poles of the vocal-tract filter. The arguments of the poles are linearly related to the formant frequencies, and the moduli of the poles are logarithmically related to the formant bandwidths. Extraction of these roots requires polynomial factorization, which has been well explored and documented through the past two centuries. However there are two reasons for avoiding this process in determining the formant frequencies. First, it requires complex arithmetic and, usually, high-precision computations. Second, the 600-bps voice digitizer does not require the formant bandwidth information. Thus the use of polynomial factorization (which provides such information) would not be fully justified unless computations are simple. Since this is not the case, a simple alternative method of estimating formant frequencies was chosen in the present 600-bps voice digitizer.

This method proceeds in two steps: first an initial approximation and then a subsequent refinement. The first step moves all the poles toward the unit circle in the z plane, so that a simple spectral analysis can provide all the formant frequencies as the initial starting point for the second step. The poles are moved toward the unit circle by simply letting the last predictive coefficient $\alpha_{n|n}$ (the product of all pole moduli) be near unity. Thus each individual pole modulus approaches unity, which implies that all the poles are near the unit circle.

If the poles move radially when $\alpha_{n|n}$ approaches unity, then the formant frequency is exact. Generally however the poles do not move radially (the ideal case) as $\alpha_{n|n}$ approaches unity; therefore the formant frequencies are shifted from their true values. These shifts do not appear to be excessive for voiced sounds. This first step produces two useful results: all formant frequencies are distinct and naturally ordered (they are separated as f_1, f_2, and f_3), and all formant frequencies are always captured. These two results are most beneficial for accomplishing successful formant tracking.

The second step is the refinement of these initial formant frequency estimates. As $\alpha_{n|n}$ moves toward it actual value, the frequency response is recomputed for a small range

around the previous formant frequency. Since $\alpha_{n|n}$ lies theoretically between -1 and 1 (in most cases somewhere between -0.5 and 0.5), a few iterations at an incremental step of -0.2 will find a sufficiently accurate formant frequency. This procedure is applied to all formant frequencies, and if any formant frequency disappears during the iteration, its previous value is retained.

When determination of the three formant frequencies is complete, the frequencies must be coded into seven bits to meet the data-rate limitation. The remaining eight bits per frame are allocated to the excitation parameters and synchronization. Normally ten bits per frame are required for coding three formant frequencies (three bits for f_1, four bits for f_2, and three bits for f_3). The most effective way of coding three formant frequencies into seven bits is by pattern matching (by coding the three formant frequencies jointly). Fortunately certain combinations of formant frequencies do not occur, a characteristic which permits a pattern-matching technique to exclude these classes in the codes. Thus the 128 formant patterns (2^7 patterns) are selected from many speech samples through a technique similar to "cluster analysis." Similarly, the six predictive coefficients are classified into 128 patterns for the unvoiced case.

At the receiver the formant frequencies are converted to six predictive coefficients and become, as in a 2400-bps LPE, the weights of the vocal-tract filter.

The subsequent sections of this report discuss the past history of formant tracking, previous 600-bps voice digitizers, and the implementation of the present 600-bps voice digitizer. A demonstration record containing samples of 600-bps speech is included with the report.

BACKGROUND

Both formant-tracking vocoders and 600-bps voice digitizers have existed for some time. This section presents some of their history. In addition the theory of linear predictive analysis is briefly reviewed, because it is the underlying principle of the present 600-bps voice digitizer.

History of Formant Tracking

The development of the formant-tracking vocoder has had a long and arduous history since its inception [1]. Its motivations were no doubt started with the publication of *Visible Speech* [2], which combined a hope of visual speech perception (for the deaf) with the successful development of the "sound spectrograph" by the Bell Telephone Laboratories. The fascinating patterns, interpreted phonemically in *Visible Speech*, combined with the apparent ease in visually identifying and tracking formants on the sound spectrograph, led to the construction of a breadboard formant vocoder by Flanagan in 1956 [3], which gave imperfect yet promising results. Flanagan's work laid the groundwork for development work during the late 1950's and early 1960's by such diverse organizations as Northeastern University (Chang [4, 5]) with the Formoder and, under government-industrial contracts, Melpar [5], Philco [6], General Dynamics (Stromberg Carlson Division)

[7], and others [8]. Most of this R&D work was supported by the government for possible military application and was largely terminated in 1966 when the government decided to use the older channel vocoders of Homer Dudley [9,10]. At this point it was recognized that the channel vocoder, although requiring at least twice the bit rate of the formant vocoder, was somewhat more intelligible and also more highly developed. A joint service effort was made to procure channel vocoders for the USC-20 program, which ultimately failed and was canceled after 4 years. But during this time the choice of a channel vocoder for this program caused further research into formant vocoders to be largely suspended.

In other parts of the world, Sir Walter Lawrence [11] sought support for the formant vocoder concept with a U.S. tour demonstrating his synthesizer PAT (parametric automatic talker) driven from formant traces, and Gunnar Fant was using his formant synthesizer OVE for basic research into speech production, leading to his book *Acoustic Theory of Speech Production* [12] in 1970. By this time it was well established that the formant vocoder concept, though attractive because its implementation would permit a lower bit rate, was frought with practical difficulties. In addition to the channel-vocoder problems of pitch tracking and voicing decision, the formant vocoder had problems of proper formant tracking, formant identification, formant acquisition for tracking after a silence, and synthesis problems, particularly in consonant production. Thus potential users of the formant vocoder became skeptical as to the probable success of this approach for low-bit-rate voice coding. This skepticism is exemplified by Moye [13], who wrote, "Although such a statement is bound to be challenged, one can say that, from the point of view of practical digital speech transmission systems, formant analysis does not work." Others [14,15] have expressed similar views.

Previous 600-bps Voice Digitizers

According to published accounts there have been at least three previous VLDR voice digitizers. By coincidence they are all 600-bps voice digitizers. Flanagan [16] demonstrated a formant-tracking vocoder operating at 600 bps and demonstrated his results in the phonograph record he attached to his article. Since the test sentence was composed of all vowels, dipthongs, and liquids ("We were away a year ago"), it was a very limited demonstration of a 600-bps voice digitizer. Nevertheless the synthesized speech was highly articulate, indicating that the formant-vocoding approach had potential for voice analysis and synthesis. Another 600-bps voice digitizer was developed by Caldwell Smith at the Air Force Cambridge Research Laboratory [17]. The device employed a pattern-matching technique to classify the channel vocoder outputs and was the result of extensive R&D work. Its intelligibility score of 92 percent for a single-talker diagnostic rhyme test (DRT) [18] was an exceptionally high score for a 600-bps voice system. A third 600-bps system consisted of a modified version of the Melpar formant vocoder, presented by tape demonstration at the 70th meeting of the Acoustical Society of America in November 1965.

Summary of Linear Predictive Analysis

Because the present 600-bps voice digitizer uses the coded output of a 2400-bps LPE, the basic principles and mathematical theory of linear predictive analysis are briefly summarized in the next few pages to facilitate discussions of the 600-bps voice digitizer. Much

of this theory has been well developed in connection with the implementation of voice digitizers operating at 2400 bps or higher rates [19-25].

In linear predictive analysis a speech sample is represented by a linear combination of past samples. Thus

$$\hat{x}_t \triangleq \alpha_{1|n} x_{t-1} + \alpha_{2|n} x_{t-2} + \ldots + \alpha_{n|n} x_{t-n}, \tag{1}$$

where x_t is a speech sample at time t, $\alpha_{j|n}$ is the jth predictive coefficient, and n is the order of prediction. A set of predictive coefficients is derived by way of minimizing the mean-square value of the prediction residual, defined by

$$\epsilon_t \triangleq x_t - \hat{x}_t. \tag{2}$$

By the application of the classical least-squares method, a set of predictive coefficients which minimizes the prediction residual, under the condition of stationarity, is obtained from

$$\begin{bmatrix} \varphi_0 & \varphi_1 & \cdots & \varphi_{n-1} \\ \varphi_1 & \varphi_0 & \cdots & \varphi_{n-2} \\ \cdots & \cdots & \cdots & \cdots \\ \varphi_{n-1} & \varphi_{n-2} & \cdots & \varphi_0 \end{bmatrix} \begin{bmatrix} \alpha_{1|n} \\ \alpha_{2|n} \\ \cdots \\ \alpha_{n|n} \end{bmatrix} = \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \cdots \\ \varphi_n \end{bmatrix} \tag{3}$$

where φ_j is the autocorrelation coefficient of the speech signal defined by

$$\varphi_j \triangleq \sum_{m=0}^{N-1-j} x_m x_{m+j}, \quad j \geq 0,$$

where N is the number of speech samples entered into the correlation analysis. Under the assumption of stationarity

$$\varphi_j = \varphi_{-j}. \tag{5}$$

Equation (3) is a set of simultaneous linear equations with a doubly symmetric coefficient matrix (a Toplitz matrix). The solution of similar problem has been encountered in statistics, and its simpler recursive solution is well known [26,27]. The solution of Eq. (3) is

$$\alpha_{i|n} = \alpha_{i|n-1} - \alpha_{n|n}\, \alpha_{n-i|n-1}, \; i = 1, 2, \ldots, n-1, \tag{6}$$

where

$$\alpha_{n|n} = \frac{\varphi_n - \sum_{i=1}^{n-1} \varphi_{n-i} \alpha_{i|n-1}}{\varphi_0 - \sum_{i=1}^{n-1} \varphi_i \alpha_{i|n-1}}, \quad n \geq 2, \tag{7}$$

and, when $n = 1$,

$$\alpha_{1|1} = \frac{\varphi_1}{\varphi_0}. \tag{8}$$

The analysis filter is a filter which generates the prediction residual as it is driven by the speech signal. Thus the analysis-filter output is the difference between the given and the predicted speech signals. Therefore the transfer function of the analysis filter, denoted by $A_n(z)$, is

$$A_n(z) = 1 - P_n(z), \tag{9}$$

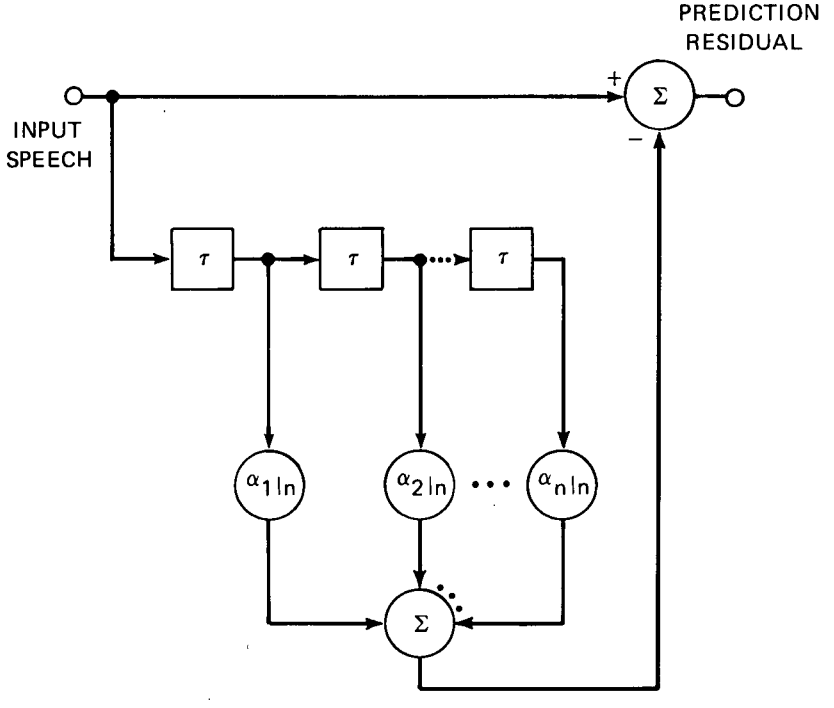
where $P_n(z)$ is the transfer function of the nth-order predictor. By the z transform of Eq. (1), $P_n(z)$ is expressed as

$$P_n(z) = \alpha_{1|n}z^{-1} + \alpha_{2|n}z^{-2} + \ldots + \alpha_{n|n}z^{-n}. \tag{10}$$

From Eqs. (9) and (10) the transfer function of the analysis filter becomes

$$A_n(z) = 1 - \left(\alpha_{1|n}z^{-1} + \alpha_{2|n}z^{-2} + \ldots + \alpha_{n|n}z^{-n}\right). \tag{11}$$

The structure of the analysis filter is shown in Fig. 3a.



(a) Analysis filter

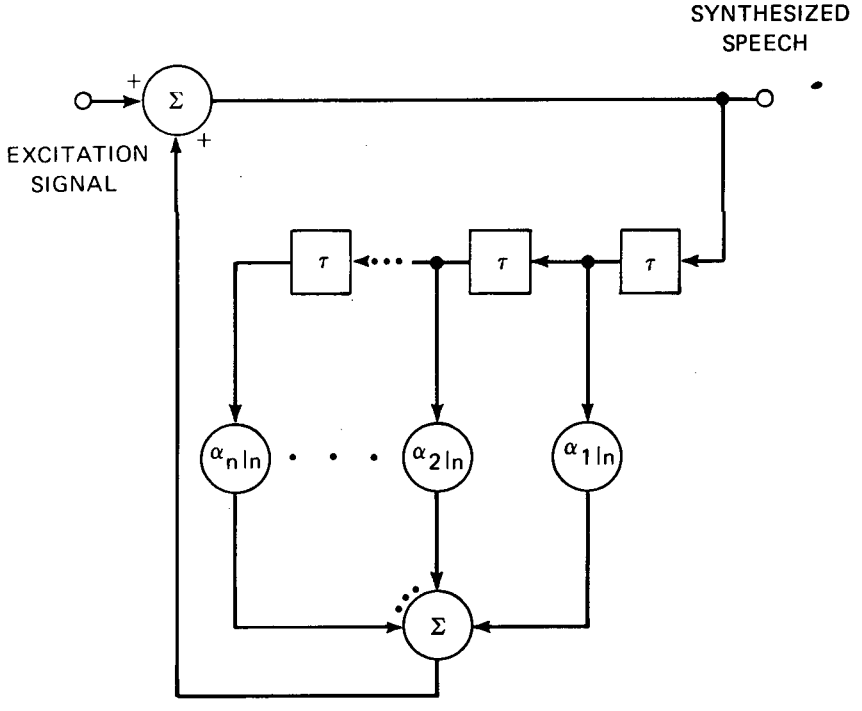
Fig. 3 — Analysis and Synthesis filters with predictive coefficients as weights

The synthesis filter (the vocal-tract filter) is an inverse of the analysis filter. Thus the transfer function of the synthesis filter, denoted by $H_n(z)$, is

$$H_n(z) = \frac{1}{A_n(z)}$$

$$= \frac{1}{1 - \alpha_{1|n}z^{-1} + \alpha_{2|n}z^{-2} + \ldots + \alpha_{n|n}z^{-n}}. \tag{12}$$

Since only the denominator of $H_n(z)$ is a function of the complex variable z, the vocal-tract filter has only poles. As a result the properties of the vocal-tract filter are entirely determined by the locations of poles. The vocal-tract filter (Fig. 3b) is structured as a positive feedback in which a predictor is in the feedback loop. If the vocal-tract filter is driven by the prediction residual, the synthesized speech would be identical to the given speech. However, a voice digitizer operating at a bit rate below the speech sampling frequency uses some form of artificial excitation.



(b) Synthesis filter

Fig. 3 (Continued) — Analysis and synthesis filters with predictive coefficients as weights

The last predictive coefficient of each iteration cycle ($\alpha_{n|n}$ expressed by Eq. (7) or (8)) is often referred to as the partial correlation coefficient, denoted by k_n:

$$k_n \triangleq \alpha_{n|n}. \tag{13}$$

It is possible to construct an analysis and a synthesis filter in which the filter weights are partial correlation coefficients. From Eqs. (6) and (11) the transfer function of the nth-order analysis filter in terms of the (n-1) th-order analysis filter is

$$A_n(z) = A_{n-1}(z) - k_n z^{-n} A_{n-1}(z^{-1}). \tag{14}$$

Let

$$B_{n-1}(z) = z^{-n} A_{n-1}(z^{-1}). \tag{15}$$

From Eqs. (14) and (15) $A_n(z)$ in terms of $A_{n-1}(z)$ and $B_{n-1}(z)$ is

$$A_n(z) = A_{n-1}(z) - k_n B_{n-1}(z). \tag{16}$$

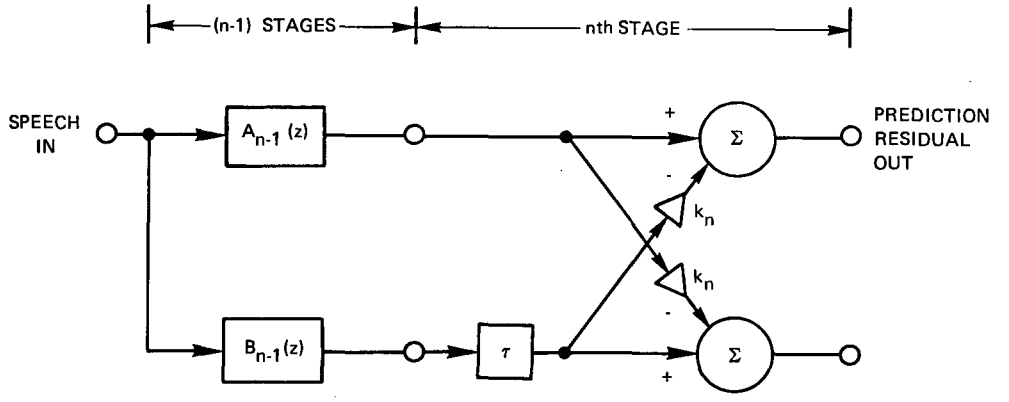
Substituting z^{-1} for z in Eq. (14) gives

$$A_n(z^{-1}) = A_{n-1}(z^{-1}) - k_n z^n A_{n-1}(z). \tag{17}$$

From Eqs. (15) and (17) $B_n(z)$ in terms of $A_{n-1}(z)$ and $B_{n-1}(z)$ is

$$B_n(z) = z^{-1}\left[B_{n-1}(z) - k_n A_{n-1}(z)\right]. \tag{18}$$

Equations (16) and (18) define the structure of the analysis filter, as shown in Fig. 4a. The performance of this cascade-lattice form of an analysis filter is identical to the transversal-filter form of an analysis filter shown in Fig. 3a.



(a) Analysis filter

Fig. 4 — Analysis and synthesis filters with partial correlation coefficients and weights

The synthesis filter is the inverse of the analysis filter. Thus its transfer function is

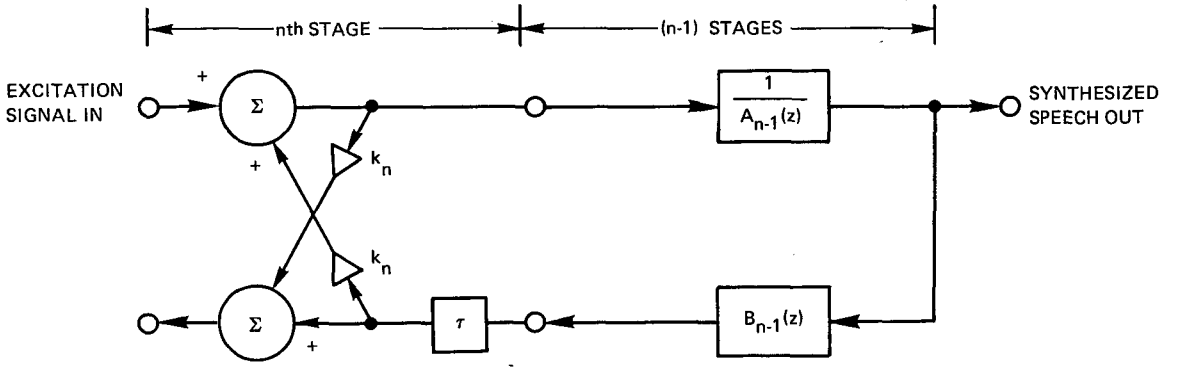
$$H_n(z) = \frac{1}{A_n(z)}. \tag{19}$$

Substitution of Eq. (16) into Eq. (19) gives

$$H_n(z) = \frac{1}{A_{n-1}(z) - k_n B_{n-1}(z)}$$

$$= \frac{\dfrac{1}{A_{n-1}(z)}}{1 - k_n \left[\dfrac{z^{-n} A_{n-1}(z^{-1})}{A_{n-1}(z)}\right]}. \tag{20}$$

Figure 4b shows the structure of the vocal-tract filter in which the filter weights are partial correlation coefficients. The filter is a cascade-lattice network. The performance of this filter is identical to that shown in Fig. 3b, provided the initial conditions for both filters are identical. As expressed by Eq. (20), the last partial correlation coefficient (k_n) behaves as the feedback gain, and the transfer function of the quantity inside the bracket is that of an all-pass filter.



(b) Synthesis filter

Fig. 4 (Continued) — Analysis and synthesis filters with partial correlation coefficients and weights

A number of significant properties of partial correlation coefficients are the following:

- The vocal-tract filter is stable if each partial correlation coefficient has a magnitude less than unity [25].

- If the vocal-tract filter is in a cascade-lattice configuration (Fig. 4b), the partial correlation coefficient can be processed directly at each filter section by the minimization of its output residual. Thus neither Eqs. (6) through (8) nor the knowledge of speech correlation coefficients ($\varphi_0, \varphi_1, \ldots$) is required. From Fig. 4a the prediction residual from the nth-stage output is expressed by

$$\epsilon_{n,t} = \epsilon_{n-1,t} - k_n \delta_{n-1,t}, \tag{21}$$

where $\epsilon_{n-1,t}$ is the output from the $A_{n-1}(z)$ filter branch (often referred to as the forward prediction residual) and $\delta_{n-1,t}$ is the output from the $B_{n-1}(z)$ filter branch (often referred to as the backward prediction residual). The partial correlation coefficient which minimizes the mean-square value of $\epsilon_{n,t}$ is

$$k_n = \frac{u_{n-1}}{P_{n-1}}, \tag{22}$$

where

$$u_{n-1} \triangleq E\left(\epsilon_{n-1,t}\delta_{n-1,t}\right) \tag{23}$$

and

$$p_{n-1} \triangleq E\left(\delta^2_{n-1,t}\right), \tag{24}$$

in which $E(\cdot)$ denotes the expectation operation which is the time-averaging process in practice. As can be noted from Eq. (22), a partial correlation coefficient is a power ratio, with the numerator being a crosscorrelation between the forward and backward prediction residuals and the denominator being the backward prediction residual power. Under the condition of stationarity the forward-prediction-residual power equals the backward prediction residual power. Equation (22) is a mathematical equivalent of the previous definition of $k_n = \alpha_{n|n}$ expressed by Eq. (7). Thus Eq. (22) is the basis for computing partial correlation coefficients directly from the analysis filter.

- The output (forward) prediction residual in terms of the input (forward) prediction residual at each section may be obtained from Eq. (21). Squaring both sides of Eq. (21) and passing the resulting quantity through the expectation operation gives

$$p_n = p_{n-1}\left(1 + k_n^2\right) - 2k_n u_{n-1}. \tag{25}$$

From Eq. (22)

$$u_{n-1} = k_n p_{n-1}. \tag{26}$$

From Eqs. (25) and (26) p_n in terms of p_{n-1} is

$$p_n = \left(1 - k_n^2\right) p_{n-1}. \tag{27}$$

- A set of partial correlation coefficients can be converted to a set of predictive coefficients by the recursion relationship expressed by Eq. (6).

- Conversely, a set of predictive coefficients can be converted to a set of partial correlation coefficients by

$$\alpha_{i|n-1} = \frac{\alpha_{i|n} + k_n\alpha_{n-i|n}}{1 - k^2}, \tag{28}$$

where $i = 1, 2, \ldots, n-1$. This relationship can be derived from the solution of the two simultaneous equations consisting of Eq. (6) and its mirror-image equation (the equation in which the index i is replaced by $n-i$).

IMPLEMENTATION OF THE 600-BPS VOICE DIGITIZER

A technical overview of the 600-bps voice digitizer was given in the Introduction. This section discusses in detail how the 2400-bps-LPE output data can be converted to 600-bps data. The items under discussion include:

- Definitions of formant frequency and formant bandwidth,
- Frequency response of the vocal-tract filter,
- A different approach to formant tracking,
- Parameter coding,
- Synthesis filter,
- Assumptions on formant synthesizer bandwidth,
- Excitation signal generation, and
- Parameter interpolation.

Definitions of Formant Frequency and Formant Bandwidth

Each pole of the vocal-tract filter may be represented by its real and imaginary parts, or its argument (the angular displacement) and modulus. The formant frequency is linearly proportional to the argument of a pole, and the formant bandwidth is logarithmically proportional to the modulus of a pole.

As was given by Eq. (12), the vocal-tract-filter transfer function is

$$H_n(z) = \frac{1}{1 - \left(\alpha_{1|n} z^{-1} + \alpha_{2|n} z^{-2} + \ldots + \alpha_{n|n} z^{-n}\right)}, \tag{12}$$

which can be rearranged as

$$H_n(z) = \frac{1}{z^{-n}\left(z^n - \alpha_{1|n} z^{n-1} - \ldots - \alpha_{n|n}\right)}$$

$$= \frac{1}{z^{-n} \prod_{i=1}^{n} (z - z_i)}, \tag{29}$$

where z_i is the ith pole of the vocal-tract filter. By the definition of the z-transform variable each pole can be expressed in terms of its real and imaginary parts. Thus

$$z_i = \exp\,[(-\mu_i + j\,\omega_i)\,\tau]$$

$$= r_i \exp\,(j\,\omega_i\,\tau), \tag{30}$$

where r_i is the radial distance of the pole z_i defined by

$$r_i = |z_i|$$

$$= \exp\,(-\mu_i\,\tau) \tag{31}$$

and τ is the sampling period of speech signals. For a stable vocal-tract filter, $\mu_i \geqq 0$, which implies that $r_i \leqq 1$.

In Eq. (30) ω_i is a formant frequency in radians per second. Solving for the formant frequency in Eq. (30) gives

$$f_i = \frac{1}{2\pi\tau} \arg\,(z_i)\ \text{Hz}. \tag{32}$$

Hence a formant frequency is linearly proportional to the argument of the corresponding pole.

The pole modulus in Eq. (31) is the envelope decay rate of the vocal-tract-filter impulse response, and μ_i is numerically equal to the real part of the ith pole in the complex s plane (the corner frequency in radians per second). Thus the 3-dB bandwidth of the ith pole, denoted by Δf_i, is

$$\Delta f_i = \frac{\mu_i}{\pi\tau}\ \text{Hz}. \tag{33}$$

From Eqs. (31) and (33) the 3-dB bandwidth in terms of the pole modulus is

$$\Delta f_i = \frac{-1}{\pi\tau}\,\ell n\,(r_i)\ \text{Hz},\ r_i \leqslant 1. \tag{34}$$

Hence, the 3-dB bandwidth of a pole is logarithmically proportional to its modulus.

Frequency Response of the Vocal-Tract Filter

The majority of previously constructed formant estimators were based on the spectral analysis of speech signals, meaning that the frequency that corresponds to a spectral envelope peak was regarded as the formant frequency. Although there are subtle differences [28], essentially similar results may be obtained by the evaluation of the vocal-tract transfer function along the unit circle in the z plane, (the frequency response).

An important point is that some of the poles may not be reflected as peaks in the frequency response of the vocal-tract filter because of their remote positions from the unit

circle and/or the interference by other poles. Hence a peak-picking process based on the vocal-tract frequency response often misses formant frequencies. Therefore certain complicated (and usually ad hoc) procedures are required to track formant frequencies [29].

Nevertheless the frequency response of the vocal-tract filter can be of value if properly used, as in the present 600-bps voice digitizer. A main advantage for using the frequency-response function is that it requires relatively simple and real arithmetics.

Since the phase response has no intrinsic value for picking peak resonance frequencies, the power-response function is used [30]:

$$W_n(z) = \frac{1}{2\pi} H_n(z)\, H_n(z^{-1}). \tag{35}$$

Substituting Eq. (12) into Eq. (35) and letting $z = \exp(j\omega\tau)$ gives

$$W(\omega) = \frac{1}{2\pi} \frac{1}{A_0 + 2\sum_{i=1}^{n} A_i \cos(i\omega\tau)}, \tag{36}$$

where

$$\left.\begin{aligned} A_0 &= 1 + \alpha_{1|n}^2 + \alpha_{2|n}^2 + \ldots + \alpha_{n|n}^2, \\ A_1 &= -\alpha_{1|n} + \alpha_{1|n}\alpha_{2|n} + \alpha_{2|n}\alpha_{3|n} + \ldots + \alpha_{n-1|n}\alpha_{n|n}, \\ A_2 &= -\alpha_{2|n} + \alpha_{1|n}\alpha_{3|n} + \alpha_{2|n}\alpha_{4|n} + \ldots + \alpha_{n-2|n}\alpha_{n|n}, \\ &\ldots, \\ A_n &= -\alpha_{n|n}. \end{aligned}\right\} \tag{37}$$

As expressed by Eq. (36), the frequency response of the vocal-tract filter is a reciprocal function of a Fourier series in which the expansion coefficients are the autocorrelation coefficients of the analysis-filter impulse response expressed by Eq. (11). The resonance frequencies of the vocal-tract filter correspond to the frequencies which make the denominator of Eq. (36) exhibit local minima. Research of local minima may be effected by the evaluation of the denominator for discretely selected frequencies. The term cos $(i\omega\tau)$ may be stored as a set of constants to facilitate the computations. If the speech sampling rate is 8000 Hz and the desired frequency resolution is 50 Hz, the term cos $(i\omega\tau)$ takes only 41 values with signs.

Figure 5 illustrates the vocal-tract-filter frequency response computed from the spoken words "happy hour." Each trace represents the spectral intensities from 0 to 4000 Hz. The trace is renewed every 25 milliseconds. As expected, unvoiced segments (/*h*/ and /*p*/) do not exhibit sharp resonance peaks, but vowels produce three to four recognizable resonance peaks. Despite the simplicity of computation the direct use of the frequency-response function does not lead to successful formant tracking.

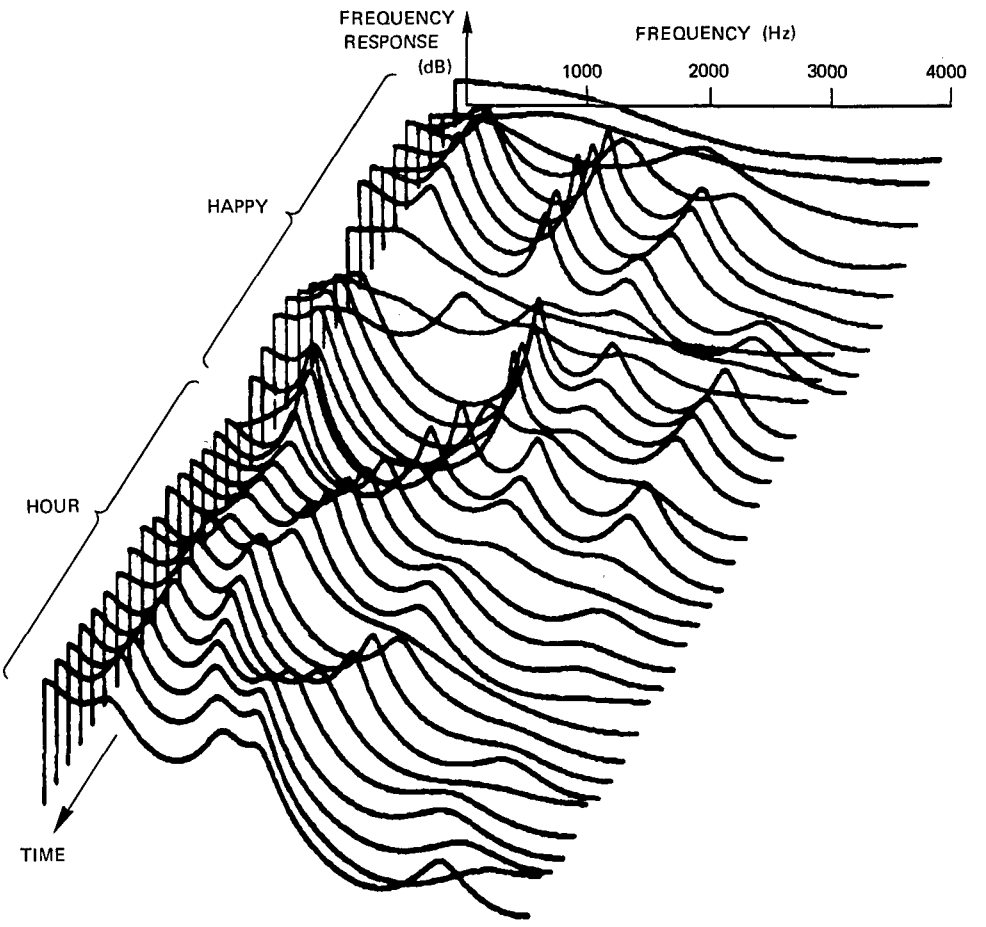


Fig. 5 — Frequency response of the vocal-tract filter estimated from actual speech signals

A Different Approach to Formant Tracking

Formant tracking is a process of estimating formant frequencies, and logging each into a designated tracker from frame to frame. Assignment of formant values to a particular track or formant number is required because each formant frequency must be interpolated during speech synthesis.

Formant tracking becomes less of a problem if the estimated formant frequencies are naturally ordered and rarely drop out. Then the lowest formant frequency simply becomes f_1, the next one becomes f_2, and so on. The present-600 bps voice digitizer requires reliable formant extraction, because it is faced with a constraint that formant frequencies are estimated only once per frame, hence the dynamics of formant frequencies during the intra-frame period are not available. Any kind of ad hoc rules or other "dead-reckoning" schemes to fill in missing formant frequencies and/or to rearrange erroneously ordered formant frequencies are virtually unworkable in practice, due to the many exceptions that arise.

The 600-bps voice digitizer employs a somewhat unconventional formant extraction method which not only provides sure acquisition but also maintains a naturally ordered formant sequence. The method proceeds in two steps: the estimation of initial (and approximate) formant frequencies, and subsequent refinements by an iterative technique.

The first step of the operation moves all the poles of the vocal-tract filter toward the unit circle in the z plane. This is accomplished by simply letting the last predictive coefficient (which is numerically equal to the last partial correlation coefficient) be near unity. Once the poles are near the unit circle, the frequency response of the vocal-tract filter exhibits extremely sharp resonance peaks. These resonance frequencies will serve as the initial iteration points to be subsequently refined by the second step of the operation. The initial resonance frequencies are approximate because the poles do not move radially as the last predictive coefficient approaches unity.

Figure 6 shows a set of vocal-tract-filter frequency responses in which the last predictive coefficient was successively varied from its actual value to near unity. The following vocal-tract-filter parameters were derived from a voiced segment of actual speech: $k_1 = 0.860$, $k_2 = -0.818$, $k_3 = -0.252$, $k_4 = 0.311$, $k_5 = 0.204$, $k_6 = 0.054$, $k_7 = 0.215$, $k_8 = -0.339$, $k_9 = 0.445$, and $k_{10} = 0.005$ (which will be varied).

In Fig. 6 the hidden second formant frequency in the original vocal-tract-filter response gradually became visible as the last predictive coefficient approached unity. This phenomena may be explained from the following three different points of view:

- *Algebraic point of view.* From Eq. (29) the poles of the vocal-tract filter in terms of the predictive coefficients are

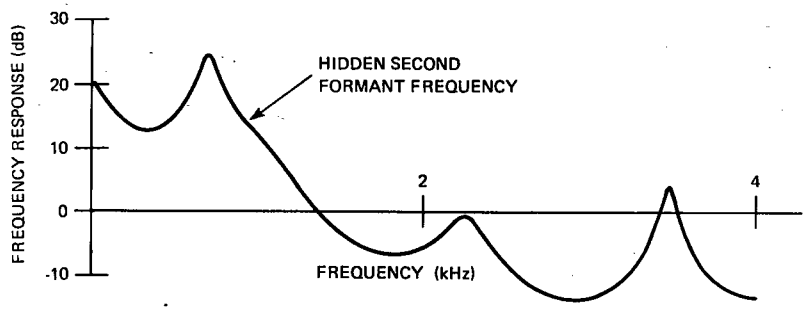
$$z^n - \alpha_{1|n} z^{n-1} - \alpha_{2|n} z^{n-2} - \ldots - \alpha_{n|n} = \prod_{i=1}^{n} (z - z_i). \tag{38}$$

Thus the last predictive coefficient ($\alpha_{n|n} = k_n$) is a product of all pole moduli of the vocal-tract filter. Therefore, by making the product be near unity, each individual pole modulus becomes near unity, signifying that all poles are near the unit circle in the z plane.

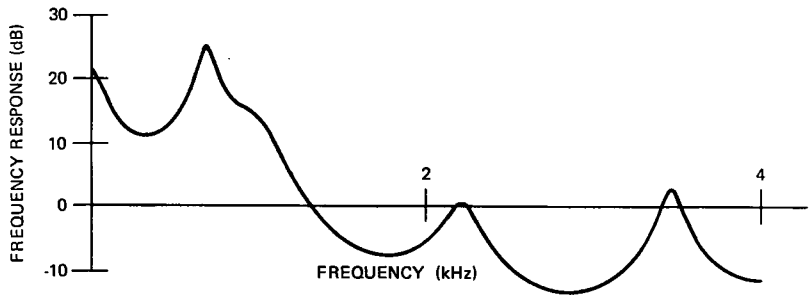
- *Control theory point of view.* The transfer function of the vocal-tract filter in terms of the partial correlation coefficients is given by Eq. (20) as

$$H_n(z) = \frac{\dfrac{1}{A_{n-1}(z)}}{1 - k_n \left[\dfrac{z^{-n} A_{n-1}(z^{-1})}{A_{n-1}(z)} \right]}, \tag{20}$$

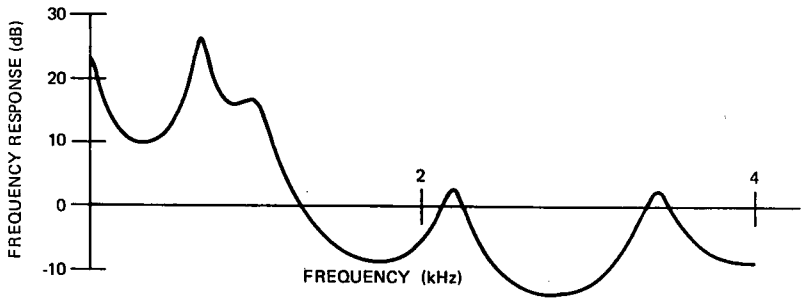
where k_n is the nth partial correlation coefficient and $A_{n-1}(z)$ is the $(n-1)$ th-order analysis-filter transfer function. The vocal-tract filter, as expressed by Eq. (20), is a positive-feedback network in which k_n behaves as a feedback gain. Since the quantity inside the bracket is a unity-gain, all-pass (frequency-independent) filter, the loop gain is determined solely by



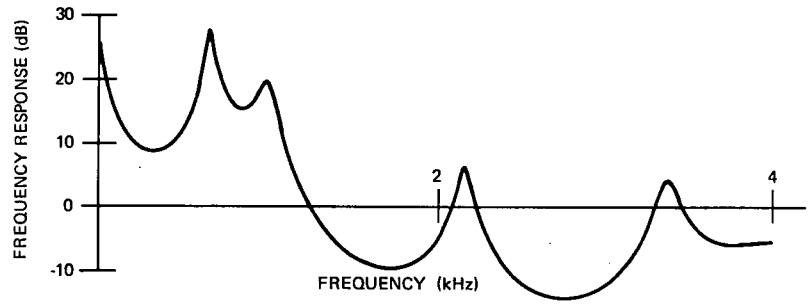
(a) Original speech with k_{10} = 0.005



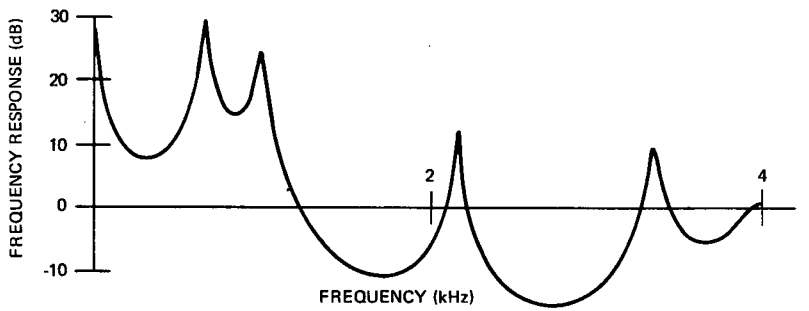
(b) With k_{10} = 0.2



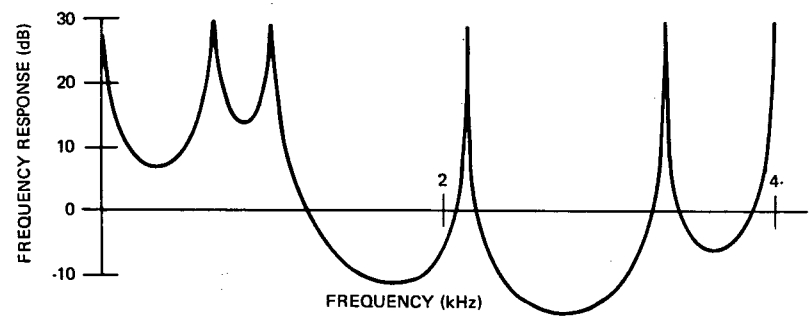
(c) With k_{10} = 0.4



(d) With k_{10} = 0.60



(e) With k_{10} = 0.80



(f) With k_{10} = 0.999

Fig. 6 — Effect of the last predictive coefficient on the frequency response of the vocal-tract filter

k_n. As k_n approaches unity, the poles migrate toward the unit circle. The trajectory of the poles as a function of the feedback gain is known as the root locus [31]. Based on the vocal-tract-filter parameters required to construct Fig. 6, the root locus is plotted, as shown in Fig. 7. As shown, the poles do not move radially, which means that the initial formant frequency estimates contain errors which are corrected by the second step of the operation.

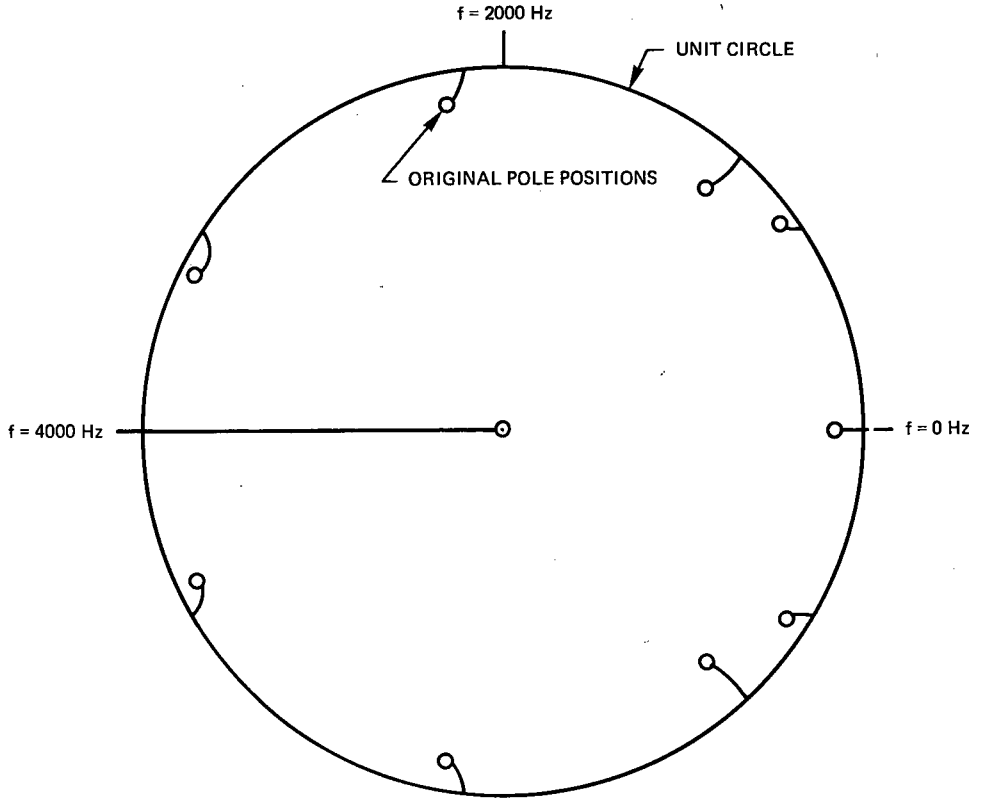


Fig. 7 — Root locus of the vocal-tract filter as the last filter coefficient approaches unity

• *Acoustic point of view.* If the effects of lung and nasal cavities are omitted, the vocal tract is closely approximated by cascaded concentric pipes, each having equal length L with different cross sectional areas $A_1, A_2, \ldots$. The reflection coefficient denoted by ρ_n is defined as the ratio of the difference to the sum of two adjacent areas. Thus

$$\rho_n = \frac{A_{n+1} - A_n}{A_{n+1} + A_n}. \tag{39}$$

It has been established that a partial correlation coefficient equals a negative value of a reflection coefficient [32]. Hence the approach of k_n to unity implies a complete reflection at one end of the vocal-tract filter (a lossless case). Thus its resonance peaks have infinitesimally small bandwidths.

Figure 8 exemplifies the effectiveness of the first step of this operation. Figure 8a is a plot of the formant frequencies derived from actual speech samples through the use of Eq. (36). As shown, the lack of continuity makes formant tracking almost impossible. Figure 8b is the result of the first step of this operation by the 600-bps voice digitizer. All formant frequencies are always present, and they are well ordered and separated!

The second step of this operation refines these initial formant estimates. As the last predictive coefficient moves toward the actual value, the frequency response is recomputed for a small range around the previous formant estimate. The theoretical range of the last coefficient is between 1 and -1. However actual speech samples show that the last coefficient is somewhere between 0.5 and -0.5. A few iterations with an incremental step of $\Delta k_n = -0.2$ will find substantially accurate formant frequencies. If a formant frequency disappears during this iteration cycle, the previous value is retained.

Parameter Coding

In a manner similar to a 2400-bps LPE, the 600-bps voice digitizer transmits two sets of speech parameters: the vocal-tract-filter parameters and the excitation parameters. The excitation parameters include the pitch period, the excitation power level, and the voice/unvoice decision. The vocal-tract-filter parameters take one of two forms depending on the voicing state: the formant frequencies for voiced sounds and the predictive coefficients for unvoiced sounds.

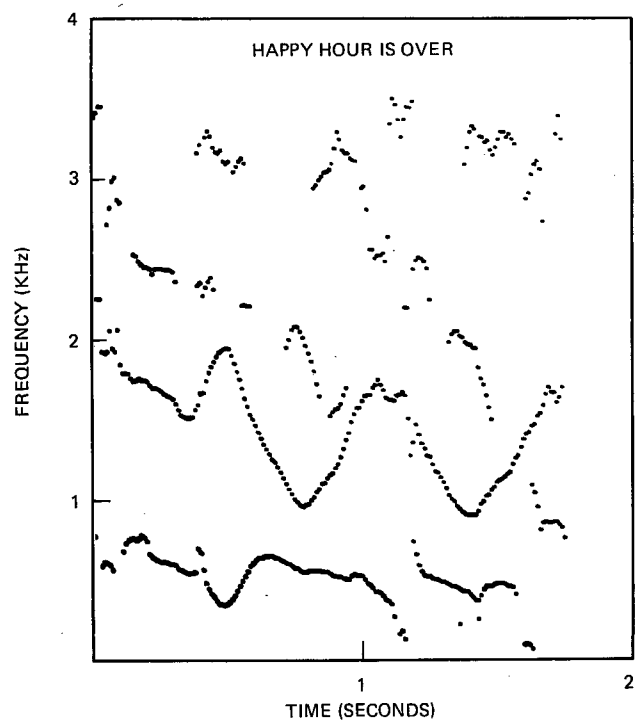
The parameter-update rate was chosen as 40 Hz, which is 10 percent slower than that of a 2400-bps LPE, due to the data-rate limitation. Thus the number of bits per frame equals 15 for this frame rate of 40 Hz.

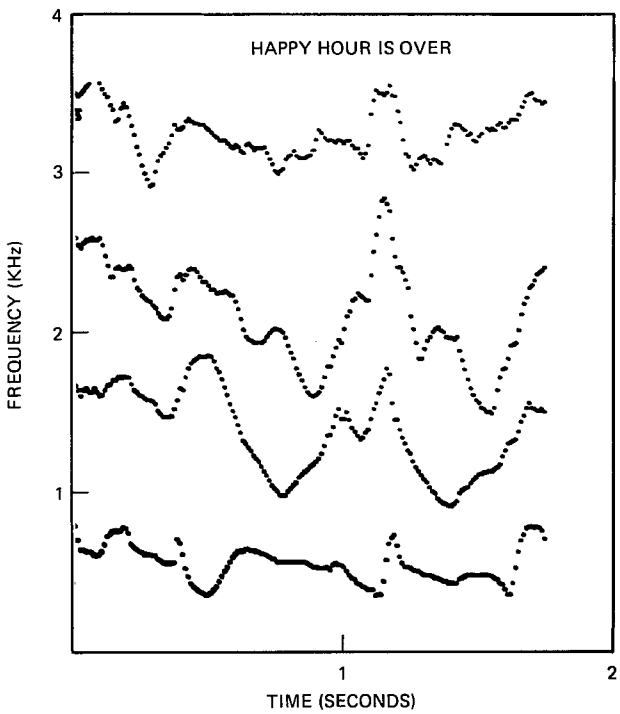
These 15 bits could be allocated in the following manner: one bit per frame for synchronization, one bit per frame for the voice/unvoice decision, four bits per frame for the amplitude information in order to encompass the dynamic range of speech encountered in normal conversation, and the last nine bits per frame for the vocal-tract-filter parameters and the pitch information. However the pitch period, even though it is a rather important parameter for the reproduction of more natural speech, possesses a contour which does not vary as rapidly as other speech parameters in normal conversation. Thus pitch information can be transmitted once every other frame without causing undue mechanical inflection in the synthesized speech, and it is quantized to five bits logarithmically from 50 to 300 Hz (12 steps per octave). The upper cutoff frequency of 300 Hz is somewhat lower than might be desired, but this is a compromise for the 600-bps voice digitizer.

Since the pitch information is transmitted once every other frame, it is necessary to group two frames in one. Therefore only one synchronization bit is required for every two frames, and the number of bits available to code vocal-tract-filter parameters becomes seven. Table 1 shows a comparison in bit assignments between a typical 2400-bps LPE and the 600-bps voice digitizer.

The vocal-tract-filter parameters control the spectral shape or tone color of the synthesized speech. A 2400-bps LPE transmits 40 bits describing the vocal-tract-filter

(a) Use of actual filter coefficients





(b) Use of actual filter coefficients except that the last coefficient is 0.9

Fig. 8 — Formant trajectory of spoken voices

Table 1 — Parameter Coding

Parameter	Coding	
	Typical 2400-Bit-Per-Second Linear Predictive Encoder	600-Bit-Per-Second Voice Digitizer
Frame rate	44.444 Hz	40 Hz
Vocal-tract-filter parameters	40 bits/frame	7 bits/frame
Excitation parameters		
Voice/unvoice decision	1 bit/frame	1 bit/frame
Amplitude	6 bits/frame	4 bits/frame
Pitch	6 bits/frame	5 bits/double frame
Synchronization	1 bit/frame	1 bit/double frame
Total number of bits	54 bits/frame	30 bits/double frame

parameters, but 600-bps voice digitizer transmits only seven bits. The reduction from 40 bits to seven bits is tantamount to a reduction from approximately 1 trillion tone colors to merely 128. Therefore the 600-bps voice digitizer must use the seven bits in the most effective way.

For voiced sounds the vocal-tract filter is well characterized by three formant frequencies. To conserve the data rate, neither formant bandwidths nor formant intensities are transmitted. On the other hand the vocal-tract filter for unvoiced sounds is poorly characterized in terms of the formant frequencies. This is because the majority of unvoiced speech spectra are broader and lack sharp resonance peaks. Consequently six partial correlation coefficients are transmitted for unvoiced sounds rather than the three formant frequencies.

At this point the question is how to code the three formant frequencies or six partial correlation coefficients so that the total number of bits per frame will not exceed seven. If each formant frequency were quantized independently, at least ten bits would be required for good speech synthesis (Table 2). Since ten bits exceeds the transmission capacity, an alternative approach was sought.

Table 2 — Formant Frequency Coding If the Formant Frequencies Were Quantized Independently

Formant Frequency	Range (Hz)	Number of Bits
f_1	150 to 1000	3
f_2	700 to 2500	4
f_3	1600 to 3100	3

In this new approach formant frequencies are not quantized independently, because they are mutually dependent (f_3 may be predicted from f_1 and f_2 for most of the vowels) and certain combinations of formant frequencies do not occur in any given language. That is, formant frequencies are highly grouped, as shown in Fig. 9 [33]. Thus a most effective coding may be achieved by the consideration of all formant frequencies jointly. This argument has led to a pattern-matching approach to formant-frequency coding.

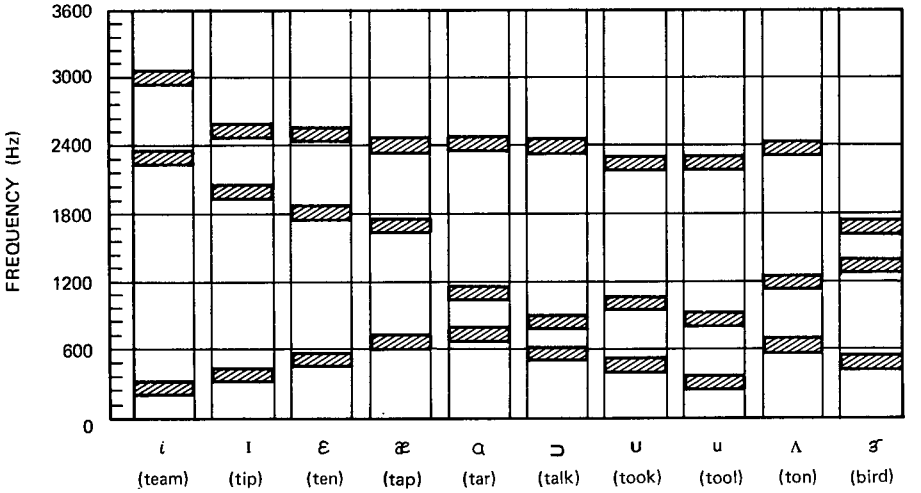


Fig. 9 — Mean formant frequencies for 33 men uttering the English vowels. (After Peterson and Barney [33].) (Words which would contain the vowel sounds are given in parentheses.)

To select 128 reference formant patterns, over 10,000 formant frequencies were collected from male and female subjects. These formant frequencies were classified into 128 patterns in such a manner that the Euclidian distance between any two reference patterns was greater than a prescribed value (R):

$$\sum_{i=1}^{3} \left[(f_{i,m} - f_{i,j})\, w_i\right]^2 > R^2,\ m, j = 1, 2, \ldots, 128,\ m \neq j, \tag{40}$$

where $f_{i,m}$ is the ith formant frequency (i = 1, 2, and 3) of the mth pattern and w_i is the weighting factor for the ith formant frequency.

These weighting factors emphasize the most important formant frequencies from a perceptual viewpoint. For example, among the first three formant frequencies, f_3 is the least important. This is demonstrated in that synthesized speech is intelligible in most cases with f_1 and f_2 only. Notable exceptions are for /r/ and /l/, which cannot reliably be distinguished by f_1 and f_2 alone. Although both f_1 and f_2 are important, it has been found that f_1 should be weighted more heavily, mainly because the level of f_1 is more

constant and errors or fluctuations in its values are more obvious to the human ear. Thus the weighting factors were chosen as $w_1 = 3$, $w_2 = 2$, and $w_3 = 1$. The magnitude of R was selected experimentally to be 400 Hz.

At the transmitter each observed formant-frequency set is compared with the stored reference formant-frequency patterns. The selected pattern is based on the minimum-distance criteria:

$$\min_m \left[\sum_{i=1}^{3} \left(f_{i,m} - \bar{f}_i \right) w_i \right]^2, \tag{41}$$

where $\bar{f}_i$ is the observed ith formant frequency. The code to be transmitted is simply the index of the chosen reference formant set.

Similar procedures are applied to classify six predictive coefficients stemming from unvoiced sounds. For unvoiced sounds description of the vocal-tract filter need not be precise. An illustration of this point is that when Fransen of NRL [34] previously applied a pattern-matching technique to classify predictive coefficients for both voiced and unvoiced sounds, the method generated high-quality speech at 1200 bps, with a diagnostic-rhyme-test (DRT) intelligibility score of 88 percent.

Synthesis filter

The synthesis filter may take many different forms: narrowband filters in series, narrowband filters in parallel, a transversal filter, or a cascade-lattice filter. Although the use of narrowband filters is simple, a cascade-lattice filter was used as the synthesizer for this system because of the following advantages:

- The transmitted vocal-tract-filter parameters for unvoiced sounds (partial correlation coefficients) can be used directly as filter weights.

- The necessary excitation power level which produces the synthesized power equal to the input speech power is obtained by a simple relationship:

$$P_{ex} = \prod_{i=1}^{n} \left(1 - k_i^2 \right) P_s, \tag{42}$$

where P_{ex} and P_s are the excitation power for the synthesizer and the input signal power respectively. Equation (42) is a direct consequence of Eq. (27).

- The intensity of the individual formant frequency is automatically weighted by the mutual locations of the poles (as in the serial analog vocal tract using narrowband filters).

- The cascade-lattice synthesis filter was already available in the Navy experimental 2400-bps LPE.

For voiced sounds the formant-frequency information must be converted to predictive coefficients. The transfer function of a filter having three pairs of complex-conjugate poles is

$$Y(z) = \frac{1}{\prod_{i=1}^{3} \left(1 - 2\epsilon^{-\mu T} \cos \omega_i T z^{-1} + \epsilon^{-2\mu T} z^{-2}\right)}, \tag{43}$$

where ω_i is the ith formant frequency in radians per second and the factor μ is related to the pole modulus as indicated by Eq. (31); and the transfer function of a sixth-order vocal-tract filter in terms of predictive coefficients is

$$H_6(z) = \frac{1}{1 - \alpha_{1|6} z^{-1} - \alpha_{2|6} z^{-2} - \ldots - \alpha_{6|6} z^{-6}}. \tag{44}$$

Comparison of Eqs. (43) and (44) term by term gives a set of predictive coefficients in terms of formant frequencies and the pole moduli (related to formant bandwidths). Thus

$$\left.\begin{aligned}
\alpha_{1|6} &= -B_1 - B_2 - B_3, \\
\alpha_{2|6} &= -r_1^2 - r_2^2 - r_3^3 - B_1B_2 - B_1B_3 - B_2B_3, \\
\alpha_{3|6} &= -(B_2 + B_3)r_1^2 - (B_1 + B_3)r_2^2 - (B_1 + B_2)r_3^2 - B_1B_2B_3, \\
\alpha_{4|6} &= -r_1^2r_2^2 - r_1^2r_3^2 - r_2^2r_3^2 - r_1^2B_2B_3 - r_2^2B_1B_3 - r_3^2B_1B_2, \\
\alpha_{5|5} &= -r_1^2r_2^2B_3 - r_1^2r_3^2B_2 - r_2^2r_3^2B_1, \\
\alpha_{6|6} &= -r_1^2r_2^2r_3^2,
\end{aligned}\right\} \tag{45}$$

where B_i is a simplified notation for

$$B_i = -2\epsilon^{-\mu T} \cos \omega_i T, \; i = 1, 2, \text{ or } 3, \tag{46}$$

and r_i, as defined by Eq. (33), is the ith-pole modulus. The relationship between $\epsilon^{-\mu T}$ and the 3-dB formant bandwidth is expressed by Eq. (34). Finally the set of predictive coefficients can be converted to a set of partial correlation coefficients through the use of Eq. (28).

Formant-Bandwidth Assumptions

Formant bandwidths depend not only on the respective formant frequencies [35] (Fig. 10) but also on the individual quality of a particular voice. However formant bandwidths are not too critical to speech intelligibility. Therefore the formant bandwidths may be approximately assigned in accordance with the formant frequencies, if the individual quality is not too important. Examples of workable assumptions are

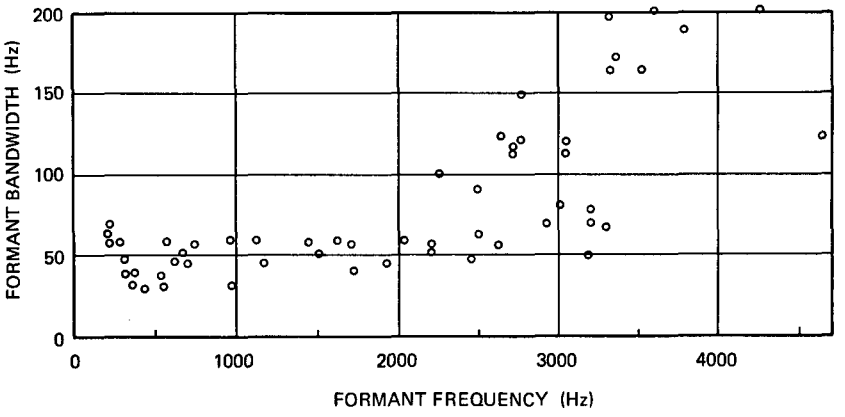


Fig. 10 — Formant bandwidth as a function of formant frequency under conditions of a closed glottis. (After Fant [35].)

$$\left.\begin{aligned} \Delta f_i &= 50 \text{ Hz, if } f_i \leqq 2000 \text{ Hz,} \\ &= 50 + 0.1\,(f_i - 2000) \text{ Hz, if } f_i > 2000 \text{ Hz,} \end{aligned}\right\} \tag{47}$$

or fixed values for each formant such as $\Delta f_1 = 50$ Hz, $\Delta f_2 = 60$ Hz, and $\Delta f_3 = 80$ Hz.

Excitation Signal Generation

The nature of the excitation signal is virtually identical to that used for a 2400-bps LPE (a pulse train for voiced sounds and random noise for unvoiced sounds). Although not mandatory, inclusion of a real pole in the pulse excitation somewhat alleviates the tendency toward a nasal quality in the synthesized voiced sounds. Likewise slightly pre-emphasized noise assists in the production of more crisp unvoiced sounds.

Parameter Interpolation

As in a 2400-bps LPE, parameters require interpolation during the intraframe period. The pitch period and unvoiced sounds require interpolation four times per frame, while the excitation power may be interpolated logarighmically pitch-synchronously for voice sounds.

The six partial correlation coefficients transmitted for unvoiced sounds need not be interpolated. On the other hand the formant frequencies transmitted for voiced sounds are interpolated pitch-synchronously. An important point is that there is not interpolation across voicing transitions, so that formant frequencies and power at the voiced onset (which is critical to the intelligibility) can be captured fully. It might be possible to further improve the initial second-formant frequency values be either retaining the previous value across unvoiced or silence intervals or using simple interpolation rules for predicting the second-formant initial value from its last known value, as has been previously suggested in the literature [36].

EXPERIMENTION

Three important tests were selected to illustrate the strengths and weaknesses of the 600-bps voice digitizer:

- The diagnostic rhyme test (DRT) of transmitted voices for the intelligibility assessment,
- The spectral analysis of synthesized speech for the visual evaluation,
- Transcription of a synthesized speech sample on a record for audition.

Intelligibility Test

An important objective of the DRT [18] is in the determination of speech perception as influenced by process parameters (the parameter update rate, the number of bits for each parameter, and the choice of parameters). The test not only provides the measure of intelligibility but also evaluates the discriminability of six distinctive features: voicing, nasality, sustention, sibilation, graveness, and compactness. The DRT word list is comprised of 448 monosyllable rhyming word pairs in which initial consonants differ by only a single feature.

Table 3 lists the DRT score of the 600-bps voice digitizer. For comparison the DRT scores of the present Navy experimental 2400-bps LPE are also listed. The DRT score of the 600-bps voice digitizer is 79.9 percent, which is an acceptable but not a particularly high score. For comparison a previous formant vocoder developed by Melpar [37] scored only 67 percent and required 1200 bps. Additional refinement of the 600-bps digitizer is in progress in the hope of improving the DRT score.

Table 3 — Summary of DRT Score at 600 bps and, For Comparison, at 2400 bps

Feature	Perception	600-bps Voice Digitizer	2400-bps LPE
Voicing	Distinguishes /b/ from /p/, /d/ from /t/, /v/ from /f/, etc.	99.9	89.6
Nasality	Distinguishes /n/ from /d/, /m/ from /b/, etc.	84.4	93.6
Sustention	Distinguishes /f/ from /p/, /b/ from /v/, /t/ from /θ/, etc.	78.1	77.0
Sibilation	Distinguishes /s/ from /θ/, /ʃ/ from /d/, etc.	60.2	93.2
Graveness	Distinguishes /p/ from /t/ /b/ from /d/, /w/ from /r/, /m/ from /n/, etc.	68.0	81.5
Compactness	Distinguishes /y/ from /w/, /g/ from /d/, /k/ from /t/, /ʃ/ from /s/, etc.	88.3	93.0
Average		79.9	88.0

Spectral Analysis of Synthesized Speech

Spectral analysis by the sound spectrograph is a simple and convenient means of evaluating the formant tracking performance of the 600-bps voice digitizer. Figure 11 shows the spectrographs of the original and synthesized speech. This example was taken from a portion of speech on the phonograph record included with this report. The sentence contains many varieties of sound elements: vowels, consonants, vowelike sounds (/r/ and /l/), a nasal sound (/n/), a voiced fricative (/ð/) and voiceless stops (/t/). In comparison with spectrograms of previous formant vocoders [Figure 10 of Ref 4] the synthesized speech of the 600-bps voice digitizer gives remarkably faithful spectral patterns.

Demonstration Record of Synthesized Speech Samples

The phonograph record included with this report contains several examples of synthesized speech at 600 bps. Each sample is composed of conversational sentences. The listener may decide as to the practicality of the 600-bps voice digitizer for voice communications from these samples. The spoken text is intentionally not given in this report, to avoid biasing the listener.

CONCLUSIONS

This report described a practical scheme whereby voice communications at a data rate of 600 bps is possible. The approach is attractive because the 600-bps voice digitizer is a simple extension of a 2400-bps linear predictive encoder which will be generally deployed by DOD and other government agencies. The 600-bps voice digitizer uses the output of the 2400-bps linear predictive encoder by converting its linear predictive coefficients to three formant frequencies and then matching the frequency patterns to preselected reference patterns for economical coded transmission.

Some speech quality is lost at 600 bps, and the synthesized speech sounds nasal, is occasionally slurred, and lacks some of the normal speaker identification capability. However the 600-bps voice digitizer can produce synthesized speech that has adequate intelligibility for specialized military voice communications. Three areas now require further investigation: improvement of the intelligibility, reduction of the prevailing nasal quality, and evaluation of the performance under transmission-error conditions.

ACKNOWLEDGMENTS

Although the work described in this report was carried out primarily by the authors, a number of persons contributed their talents and support to the total effort.

Bruce Wald, Superintendent of the NRL Communication Sciences Division, and Don I. Himes, Head of the Systems Integration and Instrumentation Branch, showed their keen interest in this work from the beginning. Their consultation and encouragement have been

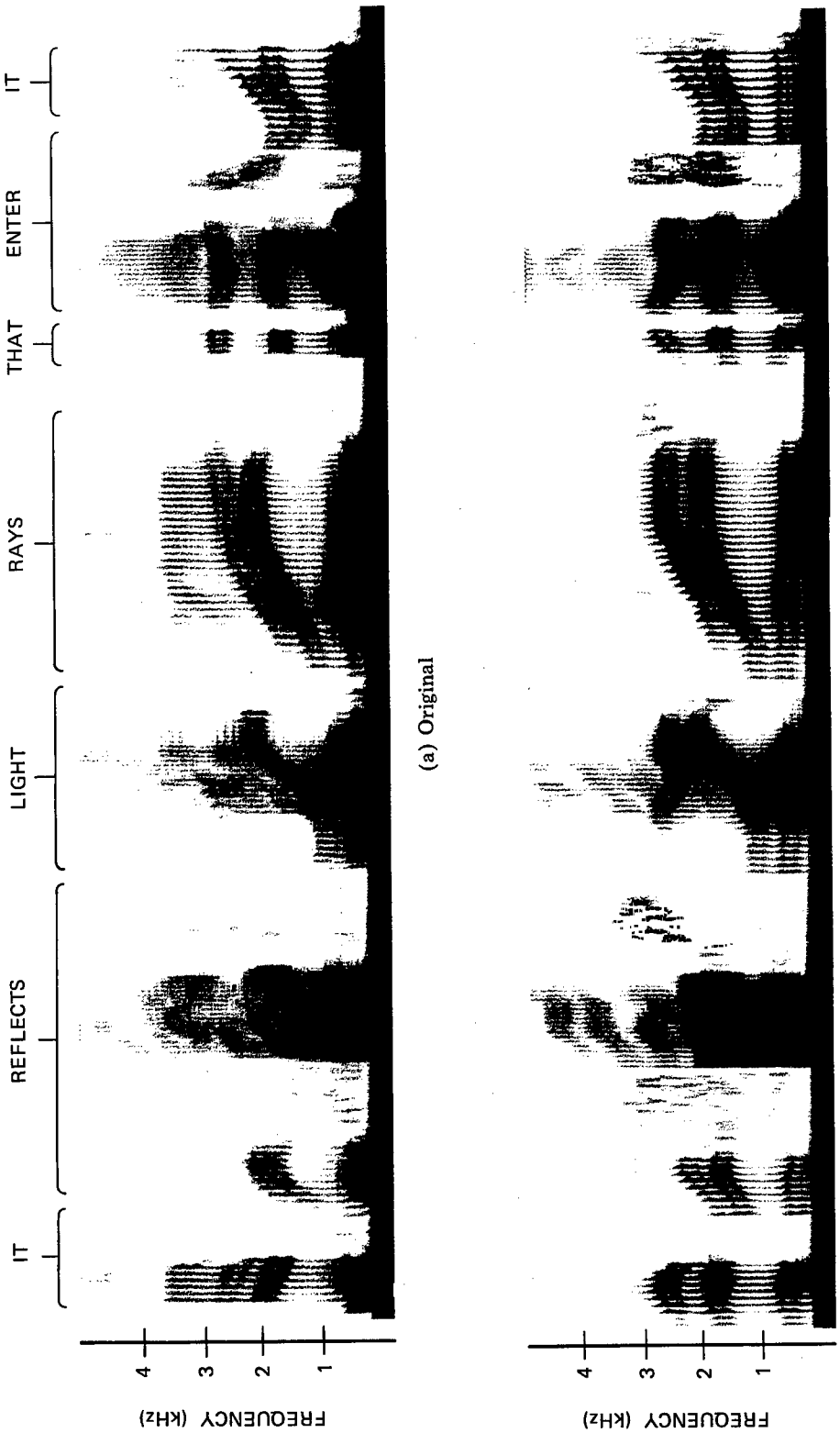


Fig. 11 — Spectral analysis of original speech and as synthesized by the 600-bps voice synthethizer

most welcome. E. Lee Kline, Head of the Integrated Systems Synthesis Section, has directed this R and D effort and coordinated NRL's work with that of other agencies having similar interests.

James Dunn of the ITT Defense Communications Division worked with the authors during the past three years and freely exchanged his ideas. Larry Fransen, a member of the NRL staff, assisted the authors by linking the Navy experimental 2400-bps LPE with the PDP 11/45 and CSP 30 computers for various experiments.

This work could not have been completed without the help of Jim West of NRL, who maintained our extensive computer facilities in good working order.

Finally, the authors gratefully acknowledge the support of Robert Martin and Irv Smietan of the Naval Electronic Systems Command, who initiated, supported, and promoted this R&D effort.

REFERENCES

1. J.C. Steinberg, "Resonance Vocoder," patent 2, 635, 146, May 1953.

2. R.K. Potter, G.A. Kopp, and H.C. Green, *Visible Speech*. van Nostrand, New York, 1947.

3. J.L. Flanagan and A.S. House, "Development and Testing of a Formant-Coding Speech Compression System," J. Acoust. Soc. Am. 28, 1099-1106 (1956).

4. S.-H. Chang, "Two Schemes of Speech Compression System," J. Acoust. Soc. Am. 28, 565-572 (1956).

5. S.J. Campanella, D.C. Coulter, et al., "Formant Tracking Vocoder System," Final Report to USAERDL (AD No. 478493), 14 Aug. 1962 issued 30 Sept. 1965.

6. D.M. Jurenko, R. Abt, and J.P. Schultz, "Speech Bandwidth Reduction," Final Report RADC (AD No. 801360), Sept. 1966.

7. F.H. Slaymaker and R.A. Houde, "Speech Compression by Analysis Synthesis," J. Audio Eng. Soc. 10, 144-148 (1962).

8. L.G. Stead and E.T. Jones, "The SRDE Speech Bandwidth Compression Project," Report 1133, Signals Research and Development Establishment, Christchurch, England, Mar. 1961.

9. H. Dudley, "Signal Transmission," U.S. Patent No. 2, 151, 091, granted 21 Mar. 1939 (filed 30 Oct. 1935).

10. H. Dudley, "Remaking Speech," J. Acoust. Soc. Am. 11, 169-177 (1939).

11. W. Lawrence, "The Synthesis of Speech From Signals Which Have a Low Information Rate," pp. 460-499 in *Communication Theory*, W. Jackson, Butterworths Sci. Publ. London, 1953.

12. G. Fant, *Acoustic Theory of Speech Production*, Monton, The Hague, 1970.

13. L.S. Moye, "Digital Transmission of Speech at Low Bit Rates," Electrical Communication (ITT) 47 (No. 4), 212-223 (1972).

14. Y. Kato and K. Ochiai, "Analysis Synthesis Telephony," The Journal of the Institute of Electronics and Communication Engineers of Japan, 51 (No. 11), 1420-1426, (Nov. 1968) (in Japanese).

15. J.L. Flanagan, *Speech Analysis, Synthesis and Perception*, Springer-Verlag, Berlin, N.Y., 1972, 2nd Ed., p. 165: "These factors conspire together to make formant analysis a difficult problem"

16. J.L. Flanagan, C.H. Coker, L.R. Rabiner, R.W. Schafer, and N. Umeda, "Synthetic Voices for Computers," IEEE Spectrum, 7 (No. 10), 22-45 (1970).

17. C.P. Smith, "Perception of Vocoder Speech Processed by Pattern Matching," J. Acoust. Soc. Am. 46 (No. 6, Part 2), 1562-1571 (1969).

18. W.D. Voiers, A.D. Sharpley, and C.J. Hehmsoth, *Research on Diagnostic Evaluation of Speech Intelligibility*, Report AFCRL-72-0694, Jan. 1973.

19. B.S. Atal and S.L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of Speech Wave," J. Acoust. Soc. Am. 50 (No. 2, Part 2), 637-655 (1971).

20. F. Itakura and S. Saito, "Digital Filtering Techniques for Speech Analysis and Synthesis," Conference Record, 7th Int. Congr. Acoustics, paper 25C1, 1971, Vol. 3, pp. 261-264.

21. E. Matsui, T. Makajima, T. Suzuki, and H. Omura, "An Adaptive Method for Speech Analysis Based on the Kalman Filtering Theory," Bull Electrotech. Lab. 36 (No. 3), 42-51 (1972) (in Japanese).

22. J.D. Markel and A.H. Gray, Jr., "On Autocorrelation Equations as Applied to Speech Analysis," IEEE Trans. Audio Electroacoust. AU-21, 69-79 (1973).

23. J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE 63 (No. 4), 561-580 (1975).

24. J.L. Melsa, A.P. Sage, et al., *Development of a Configuration Concept of a Speech Digitizer Based on Adaptive Estimation Techniques*, Prepared for Defense Communication Agency, Southern Methodist University, 329 pages, 1973.

25. G.S. Kang, *Application of Linear Prediction Encoding to a Narrowband Voice Digitizer*, NRL Report 7774, Oct. 1974.

26. J. Durbin, "The Fitting of Time-Series Model," Rev. Int. Inst. Stat. 28, 233-244 (1960).

27. N. Levinson, "The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction," J. Math. Phys. 25 (No. 4), 261-278 (1947).

28. H.F. Silverman and N.R. Dixon, "A Parametrically Controlled Spectral Analysis System for Speech," IEEE Trans. on Acoustics, Speech and Signal Processing ASSP-22 (No. 5) 362-381 (1974).

29. J.D. Markel, "Formant Trajectory Estimation From a Linear Least-Squares Inverse Filter Formulation," SCRL Monograph 7, 1971.

30. F. Itakura and S. Saito, "A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies," Electronics and Communications in Japan 53-A (No. 1), 36-43 (1970).

31. J.T. Tou, "Digital and Sampled-Data Control Systems," Section 2.6, McGraw-Hill, New York, 1959.

32. H. Wakita, "Estimation of the Vocal Tract Shape by Optimal Inverse Filtering and Acoustic/Articulatory Conversion Methods," SCRL Monograph 9, 1972.

33. G.E. Peterson and H.L. Barney, "Control Methods Used in a Study of the Vowels," J. Acoust. Soc. Am. 24, 175-184 (1952).

34. L.V. Fransen, *Application of Pattern Matching to Linear Predictive Coding of Speech at* 1200 *bps*, NRL Report 7931 Oct. 1975.

35. G. Fant, *Speech Sounds and Features*, The MIT Press, 1973, p. 9.

36. J.M. Pickett and D.C. Coulter, "Statistics of F2 Adjacent to Consonants and Prediction of F2 Onsets," J. Acoust. Soc. Am. 39, 953-959 (1966).

37. J.W. Hale, S.F. Muzidal, and W.J. Richter, Jr., "*Narrow Band Analoy Modem*," Final Report on ONR Contract N00014-69-C-0102, Jan. 1970.

BIBLIOGRAPHY ON FORMANT ANALYSIS AND/OR SYNTHESIS

B.P. Bogert, "On the Band Width of Vowel Formants," J. Acoust. Soc. Am. 25 (No. 4), 791-792 (1953).

D.C. Coulter, "System for Determining Consonant Formant Loci," Patent 3, 268, 662, 1966.

C.G.M. Fant, J. Martony, et al., "Recent Progress in Formant Synthesis of Connected Speech," J. Acoust. Soc. Am. 33 (No. 6), 834-835 (1961) (abstract).

G. Fant and J. Martony, "Quantization of Formant Coded Synthesis Speech," STL-QPSR 2, 16-18, 1961.

G. Fant and A. Risberg, "Auditory Matching of Vowels With Two Formant Synthetic Sounds," STL-QPSR 4, 7-11, 1963.

G. Fant, *Speech Sounds and Features*, The MIT Press, 1973, Chapter 5.

J.L. Flanagan, "A Speech Analyzer for a Formant-Coding Compression System," MIT Acoustics Lab., Sci. Rep. 4, USAF Contract AF19(604)-626, 1955.

J.L. Flanagan, "Automatic Extraction of Formant Frequencies From Continuous Speech," JASA, Vol. 28, No. 1, 110-118, 1956.

J.L. Flanagan, "Evaluation of Two Formant-Extraction Devices," J. Acoust. Soc. Am., 28 (No. 1), 118-125 (1956).

J.L. Flanagan, "Band Width and Channel Capacity Necessary to Transmit the Formant Information of Speech," J. Acoust. Soc. Am. 28 (No. 4), 592-596 (1956).

J.L. Flanagan, C.H. Coker, and C.M. Bird, "Digital Computer Simulation of Formant-Vocoder Speech Synthesizer," Audio Eng. Soc. Meeting, 1963.

J.L. Flanagan, *Speech Analysis, Synthesis and Perception*, Springer-Verlag, New York, 1972, Section 8.5.

B. Gold and L.R. Rabiner, "Analysis of Digital and Analog Formant Synthesizers," IEEE Trans. on Audio and Electroacoustics AU-16 (No. 1) 81-94 (1968).

A.S. House and K.N. Stevens, "Estimation of Formant Bandwidths From Measurements of Transient Response of the Vocal Tract," J. Speech and Hearing Research 1 (no. 4), 309-315 (1958).

C.R. Howard, "Speech Analysis-Synthesis Scheme Using Continuous Parameters," J. Acoust. Soc. Am. 28 (No. 6), 1091-1098, 1956.

J. Martony and G. Fant, "Information Bearing Aspects of Formant Amplitude," pp. 409-411 in Proc. 5th Int. Congr. Phon. Sci., Münster, 1964.

L.R. Rabiner, "Digital Formant Synthesizer for Speech-Synthesis Studies," J. Acoust. Soc. Am. 43 (No. 4), 822-828 (1968).

L.R. Rabiner, R.W. Schafer, and J.L. Flanagan, "Computer Synthesis of Speech by Concatenation of Formant-Coded Words," Bell Sys. Tech. J. 50 (No. 5), 1541-1558 (1971).

L.R. Rabiner and R.W. Shafer, "Speech Analyzer-Synthesizer System Employing Improved Formant Extractor," Patent 3, 649, 765, Mar. 1972.

T.J. Zebo and W.C. Lin, "On the Accuracy of Formant Parameter Estimation Based on the Method of Prony," 1972 Conf. on Speech Communication and Processing, pp. 85-88 in Conference Record, AFCRL Data Sciences Laboratory, Bedford, Mass., 1972.